Definition and Simple Examples

Substitution Estimator for Causal Parameters: ATE

G-Comp Estimation of the ATE

Non-parametric Boostrap

Inverse Propensity of

G-Computation and IPTW

SSE 708: Machine Learning in the Era of Big Data

David McCoy PhD, MSc
Division of Biostatistics, UC Berkeley
Edward's Lifesciences

July 7-11 2025

Table of Contents

1. Definition and Simple Examples
2. Substitution Estimator for Causal Parameters: ATE
3. G-Comp Estimation of the ATE
4. Non-parametric Boostrap
5. Inverse Propensity of Treatment Weighting
6. The Influence Function
7. IPTW Coding: Estimating the ATE in HEARTFIX with IPTW
8. Propensity Shift Example
9. Summary

Definition and Simple Examples

Substitution Estimator for Causal Parameters: ATE

G-Comp Estimation of the ATE

Non-parametric Boostrap

Inverse Propensity of

What is a substitution estimator?

- If the parameter of interest is some algorithm (mapping), Ψ applied to the true data-generating distribution, P_0, so $\Psi(P_0)$ equals the quantity of interest, a substitution substitutes an estimate of P_0, say P_n^* and then uses the same mapping, or $\Psi(P_n^*)$.
- For example, consider a simple situation with $O = Y \sim P_0$, $\Psi(P_0) = E_0(Y) = \sum_y y * P_0(Y = y)$.
- Then, $\Psi(P_n) = \sum_y y * P_n(Y = y)$ is the substitution estimator.
- Assume the data are n independent observations of $Y_i, i = 1, ..., n$.
- The empirical distribution P_n just assigns probability $1/n$ to every observation, so the substitution estimator can be rewritten as:

$$\Psi(P_n) = \frac{1}{n} \sum_{i=1}^{n} Y_i = \bar{Y}$$

or just the sample average.

Example: Sample Variance

- Setup

$$\begin{aligned} Data &= Y_i, i = 1, ..., n, \\ \Psi(P_0) &= var_0(Y) = E_0(Y - E_0)^2 \\ &= \sum_y (y - E_0 Y)^2 P_0(Y = y) \end{aligned}$$

- We derived the substitution estimator for the mean $E_0(Y)$ is $\bar{Y}$
- Again, we plug in P_n for the unknown P_0 and get

$$\begin{aligned} \Psi(P_n) &= var_n(Y) = \\ &= \sum_y (y - \bar{Y})^2 P_n(Y = y) \\ &= \frac{1}{n} \sum_{i=1}^{n} (Y_i - \bar{Y})^2 \end{aligned}$$

Substitution Estimator for Average Treatment Effect (ATE), $\Psi(P_X) = (Y_1 - Y_0)$

[Robins(1999)] proposed substitution estimators (G-computation) for causally inspired estimands.

- Recall that under assumptions:

$$E(Y(1)-Y(0)) = E_{W,0}\{E_0(Y|A=1,W)-E_0(Y|A=0,W)\}$$

  → Data is: $O = (W, A, Y) \sim P_0 \in \mathcal{M}^{NP}$

  → Under assumptions discussed earlier, $\Psi(P_X) = \Psi(P_0) = \Psi(Q_0)$, where Q_0 represents both the distribution of $Y \mid W, A$ and distribution of W.

$$\Psi(Q_0) = E_{W,0}\{E_0(Y \mid A=1, W) - E_0(Y \mid A=0, W)\}$$

  → Let $Q_0(A, W) \equiv E_0(Y \mid A, W)$ and $Q_{0,W}(w) = P_0(W = w)$, then

$$\Psi(Q_0) = \sum_w \{Q_0(1, w) - Q_0(0, w)\} Q_{0,W}(w)$$

Substitution Estimator for ATE

- The estimand (parameter) of interest is:

$$\Psi(Q_0) = \sum_w \{Q_0(1, w) - Q_0(0, w)\} Q_{0,W}(w)$$

- The Substitution Estimator

$$\Psi(Q_n) = \frac{1}{n} \sum_{i=1}^{n} \{Q_n(1, W_i) - Q_n(0, W_i)\}$$

  → $Q_{n,W}(W_i) = 1/n$ (the empirical) and $Q_n(A, W)$ is a regression (or machine learning) regression of Y on (A, W).

- Provides a general approach for nonparametric estimation of parameters using machine learning.

Another Example: Counterfactual Mean Difference within Subgroups

David McCoy PhD, MSc Division of Biostatistics, UC Berkeley Edward's Lifesciences

Definition and Simple Examples

Substitution Estimator for Causal Parameters: ATE

G-Comp Estimation of the ATE

Non-parametric Boostrap

Inverse Propensity of

- Data is like above ($O = (W, A, Y) \sim P_0 \in \mathcal{M}^{NP}$).
- Additionally define a variable V which is one of the W's, so $V \subset W$.
- Causal Parameter is $\Psi(P_X)(v) = E_X(Y_1 - Y_0 \mid V = v)$.
- Say V is categorical age, then the parameter above is the stratified average treatment effect, within strata of age $V = v$.
- Estimand With assumptions, then $\Psi(P_X)(v) =$

$$\Psi(P_0)(v) = E_0\{E_0(Y \mid A = 1, W) - E_0(Y \mid A = 0, W) \mid V = v\}$$

- Substitution Estimator

$$\Psi(Q_n)(v) = \frac{1}{n_v} \sum_{i=1}^{n_v} I(V_i = v)\{Q_n(1, W_i) - Q_n(0, W_i)\}$$

where n_v is the number of observations with $V_i = v$.

Understanding concepts in statistics from simulation

- One of the best methods to understand:
  1. how causal graphs are connected to data,
  2. understanding what the target estimand (parameter of interest)
  3. how to estimate from data the parameter of interest, and
  4. what statistical measures of uncertainty mean (the sampling distribution).
- We will now use this tool to solidify the understanding thus far of the topics we've been discussing (defining parameter of interest, data-generating mechanism, parameter of interest, estimation and inference).

The HEARTFIX Trial: A Realistic Scenario

Study Design: Observational study of heart failure patients

Population: 1,000 heart failure patients from multiple hospitals

- Age 30-90 years (mean 68)
- NYHA Class II-IV symptoms
- Reduced ejection fraction
- Various comorbidities

Treatment: New heart failure device (vs. standard medical therapy)

Outcome: Heart failure hospitalization within 6 months

The Challenge: Treatment assignment wasn't randomized!

The Research Question

David McCoy PhD, MSc Division of Biostatistics, UC Berkeley Edward's Lifesciences

Definition and Simple Examples

Substitution Estimator for Causal Parameters: ATE

G-Comp Estimation of the ATE

Non-parametric Boostrap

Inverse Propensity of

What is the causal effect of the new device on heart failure hospitalizations?

Specifically, we want to estimate:

Average Treatment Effect (ATE)

$$\text{ATE} = E[Y^1] - E[Y^0]$$

$$= P(\text{hospitalization if everyone got device}) -$$

$$P(\text{hospitalization if no one got device})$$

Problem: We can't observe both Y^1 and Y^0 for the same patient!

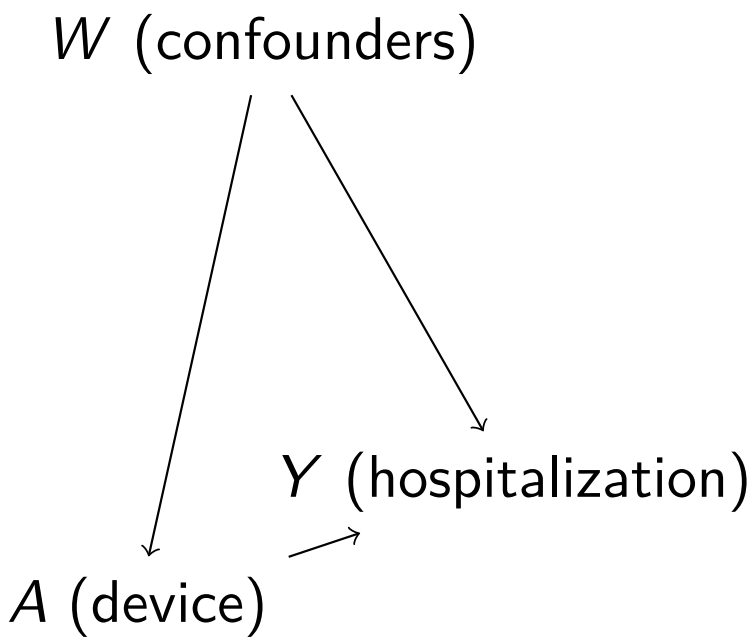
The Confounding Problem

In reality: Doctors didn't randomly assign the device...

More likely to get device:

- Younger patients
- Higher NYHA class (sicker)
- Lower ejection fraction
- Higher biomarkers (BNP)
- More comorbidities
- Prior hospitalizations

Confounding diagram:



Backdoor path: $A \leftarrow W \rightarrow Y$

Bottom Line: Simple comparison of treated vs. untreated will be **biased**!

Why Naive Analysis Fails

Naive approach: Compare hospitalization rates between groups

$$\hat{\text{ATE}}_{\text{naive}} = \bar{Y}_{A=1} - \bar{Y}_{A=0}$$

What's wrong with this?

- Device group: sicker patients at baseline
- Control group: healthier patients at baseline
- Difference in outcomes = treatment effect + selection bias

Mathematical Problem

$$E[\bar{Y}_{A=1} - \bar{Y}_{A=0}] \neq E[Y^1 - Y^0]$$

The naive estimator is **not** the causal parameter we want!

Enter G-Computation

Key Insight: If we can model $E[Y|A, W]$ well, we can "standardize" over the confounder distribution

G-Computation Algorithm:

1. Fit outcome model: $\hat{Q}(A, W) = \hat{E}[Y|A, W]$
2. Predict for everyone under $A = 1$: $\hat{Q}(1, W_i)$
3. Predict for everyone under $A = 0$: $\hat{Q}(0, W_i)$
4. Take average difference: $\frac{1}{n} \sum_i [\hat{Q}(1, W_i) - \hat{Q}(0, W_i)]$

The Magic

This gives us the **standardized** effect - what we'd see if treatment were randomly assigned in our population!

Why SuperLearner?

Challenge: We need to model $E[Y|A, W]$ accurately

Traditional approach: Assume a parametric model (e.g., logistic regression)

$$\text{logit}(E[Y|A, W]) = \beta_0 + \beta_1 A + \beta_2 W_1 + \ldots$$

Problem: What if the model is wrong?

SuperLearner solution:

- Try multiple algorithms: GLM, Random Forest, etc.
- Use cross-validation to find optimal weights
- Create ensemble that minimizes prediction error
- More robust to model misspecification

Result: Better outcome predictions $\Rightarrow$ better causal estimates!

The Uncertainty Problem

We have a point estimate, but how certain are we?

Challenges with G-computation + SuperLearner:

- Complex, non-linear algorithms
- No simple formula for standard errors
- Cross-validation adds complexity

Bootstrap to the rescue!

1. Resample data with replacement (B times)
2. Run G-computation on each bootstrap sample
3. Use distribution of estimates for inference

Bootstrap Confidence Interval

$$95\%\ \text{CI} = [2.5\text{th percentile}, 97.5\text{th percentile}]$$

Non-Parametric Bootstrap: *How it Works*

1. **Treat the observed data as "the population."**
   The n rows in your dataset—and only those rows—define the empirical distribution we will sample from.
2. **Draw many new datasets.**
   For each bootstrap replicate $b = 1, \ldots, B$
   - sample n rows *with replacement* from the original data;
   - fit the same estimator to this resampled data and store the result $\widehat{\psi}^{(b)}$.
3. **Use the collection of results as a stand-in** for the unknown sampling distribution of your estimator. The histogram of $\{\widehat{\psi}^{(1)}, \ldots, \widehat{\psi}^{(B)}\}$ mimics "what would happen" if we could repeat the whole study.
4. **Read off whatever you need:**
   - the standard error is the standard deviation of the bootstrap estimates;

Bootstrap: *Why* it Works (Intuition)

- → **Plug-in logic.** If the estimator behaves smoothly, replacing the true distribution with the empirical one barely changes its math. So resampling from the empirical data mimics sampling from the real world.
- → **Large-sample guarantee.** As the sample size grows, the distribution of the bootstrap estimates—*conditional on the observed data*—converges to the true sampling distribution. (Efron 1979; Efron Tibshirani 1993)
- → **Model-free.** No parametric assumptions are needed; the bootstrap automatically carries forward any skewness, heteroskedasticity, or other quirks present in the data.
- → **Caveats.** It may fail for very irregular estimators (e.g., medians with heavy ties) or when observations are highly dependent (then use block or cluster bootstrap variants).

What We'll Do in the Lab

Part 1: Generate realistic heart failure data

- Patient characteristics with realistic confounding
- Treatment assignment depends on patient severity
- Outcome depends on both treatment and confounders

Part 2: Calculate the "truth" (oracle knowledge)

- We know the data-generating process
- Can calculate exact causal effect
- Use this to evaluate our methods

Part 3: Implement G-computation with SuperLearner

Part 4: Bootstrap confidence intervals

Part 5: Coverage assessment - does our 95% CI actually cover 95% of the time?

Learning Objectives

By the end of this lab, you will be able to:

1. **Recognize confounding** in observational data and explain why naive analysis fails
2. **Implement G-computation** using SuperLearner for flexible outcome modeling
3. **Apply bootstrap methods** to get valid confidence intervals for complex estimators
4. **Assess method performance** by comparing estimates to known truth
5. **Evaluate coverage properties** of confidence interval procedures

Let's Code It

Load the g_comp.R file in Rstudio and let's look it over.

Residual (Un-measured) Confounding

→ **What?** Missing or badly-measured covariates that affect both treatment and outcome.

→ **Effect on G-comp.**

- Outcome model $Q_n(A, W)$ can't adjust bias stays even with huge n.
- CIs may look "tight" yet miss the truth.

→ **Red flags.** Estimate drifts toward crude risk difference; sensitivity checks unstable.

→ **Mitigate.** Collect richer W, use negative controls, or bias-analysis.

Positivity / Overlap Violations

→ **Positivity.** Every covariate pattern must allow both A=0 and A=1 with non-zero chance.

→ **When broken:**

- Propensity scores ≈ 0 or 1; deterministic treatment pockets.

→ **Impact on G-comp.**

- Must extrapolate where no data high model dependence, inflated variance.

→ **What to do.**

- Diagnose with PS histograms; restrict to common support.
- Report overlap diagnostics; consider doubly-robust estimators (TMLE/IPTW hybrid).

Inverse Probability of Treatment Weighting (IPTW)

→ **Goal.** Recover the same counterfactual means as G-comp, e.g. $[Y_1]$ and $[Y_0]$, *without* modelling $Y \mid A, W$.

→ **Idea.** Re-weight the data to form a *pseudo-population* in which treatment is independent of covariates:

$$\text{weight}_i = \begin{cases} \dfrac{1}{\hat{p}(A_i = 1 \mid W_i)}, & A_i = 1 \\ \dfrac{1}{1 - \hat{p}(A_i = 1 \mid W_i)}, & A_i = 0 \end{cases}$$

where $\hat{p}(A \mid W)$ can be learned flexibly (e.g. SuperLearner).

→ **Estimator.**

$$\hat{\psi}_{\text{IPTW}} = \underbrace{\frac{\sum_i w_i A_i Y_i}{\sum_i w_i A_i}}_{\text{weighted } Y_1} - \underbrace{\frac{\sum_i w_i (1 - A_i) Y_i}{\sum_i w_i (1 - A_i)}}_{\text{weighted } Y_0}$$

Why IPTW Makes Sense (Graphical View)

G-Computation and IPTW

David McCoy PhD, MSc Division of Biostatistics, UC Berkeley Edward's Lifesciences

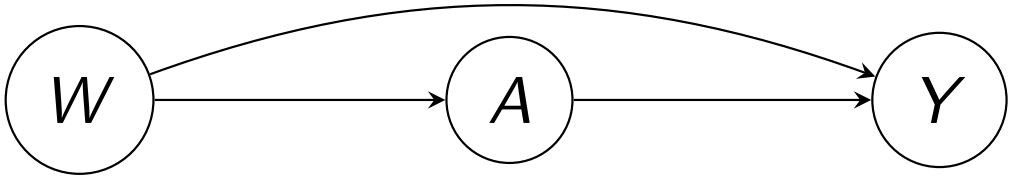
Definition and Simple Examples

Substitution Estimator for Causal Parameters: ATE

G-Comp Estimation of the ATE

Non-parametric Boostrap

Inverse Propensity of



Reweight so $A \perp\!\!\!\perp W$

Result: in the *pseudo-population*, treatment is independent of covariates, so a simple weighted mean of Y by A identifies the same counterfactual means as G-comp.

Why IPTW Makes Sense

Likelihood Factorisation

$$P(O) = P(Y \mid A, W)\, P(A \mid W)\, P(W)$$

- G-comp tackles $P(Y \mid A, W)$ directly.
- **IPTW** instead cancels out $P(A \mid W)$ by *dividing* the likelihood with the estimated propensity score—hence the weights.
- Resulting weighted sample behaves as if treatment were random: covariate balance $\Rightarrow$ unbiased plug-in mean of Y for each A level.
- **Key intuition:**
  Give more importance to people who received an unlikely treatment for their profile—they carry information about the missing counterfactual world.

Why IPTW ⇒ Influence Function, not Bootstrap

- **G-comp recap**: we bootstrapped because $\widehat{\Psi}_{\text{G-comp}}$ is a black-box plug-in.
- **IPTW idea**: build a pseudo-population with weights $w_i = 1/\widehat{p}(A_i \mid W_i)$ for treated, and $w_i = 1/\{1 - \widehat{p}(A_i \mid W_i)\}$ for controls.
- **Estimator:**

$$\hat{\psi}_{\text{IPTW}} = \underbrace{\frac{\sum_i w_i A_i Y_i}{\sum_i w_i A_i}}_{\text{weighted mean } Y \text{ in } A=1} - \underbrace{\frac{\sum_i w_i (1 - A_i) Y_i}{\sum_i w_i (1 - A_i)}}_{\text{weighted mean } Y \text{ in } A=0}$$

- **Take-home**: one pass yields the point estimate *and* its analytic SE via the influence function—no 500× resampling.

What is an Influence Function?

Core concept: How much does each observation affect our estimate

Simple example: Sample mean

- Mean height = 170 cm (n=100)
- Remove one person (190 cm) $\rightarrow$ new mean = 169.8 cm
- That person's influence = 0.2 cm

In causal inference:

- Some patients have extreme weights (IPTW)
- Influence function quantifies each person's contribution
- Identifies if few observations drive results

Key property: Influences sum to zero but variance matters

Sampling Distribution: One Study vs. Many Studies

David McCoy PhD, MSc Division of Biostatistics, UC Berkeley Edward's Lifesciences

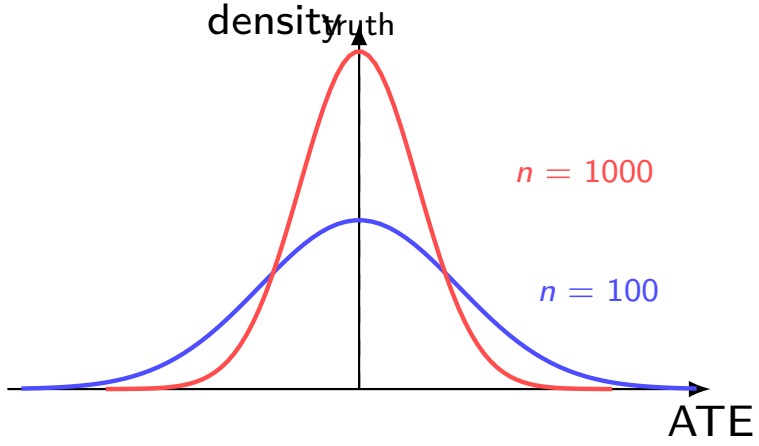
Definition and Simple Examples

Substitution Estimator for Causal Parameters: ATE

G-Comp Estimation of the ATE

Non-parametric Boostrap

Inverse Propensity of



- **Thought experiment**: repeat HeartFix many times $\Rightarrow$ each run gives an ATE.
- Those repeats form the sampling distribution.
- Larger $n \Rightarrow$ bell curve narrows (smaller SE).
- Hinges on *exchangeability* + *positivity* + a consistent estimator.

Why Estimates Vary—Even with a "Perfect" Estimator

David McCoy PhD, MSc Division of Biostatistics, UC Berkeley Edward's Lifesciences

Definition and Simple Examples

Substitution Estimator for Causal Parameters: ATE

G-Comp Estimation of the ATE

Non-parametric Boostrap

Inverse Propensity of

- Each experiment draws a new random sample $\{(W_i, A_i, Y_i)\}_{i=1}^{n}$ from the same population distribution P_0.
- The sample's mix of covariates W shifts the realised treatment probabilities $A=1$ a little bit: some runs over-represent severe patients, others mild.
- **Result**: $\hat{\psi}$ bounces around the truth. With i.i.d. data and finite variance

$$\sqrt{n}\,(\hat{\psi} - \psi_0) \xrightarrow{d} N\Big(0,\ \sigma_{\mathrm{IF}}^2\Big),$$

  a direct corollary of the Central Limit Theorem applied to the influence-function representation.
- σ_{IF}^2 is exactly "how much one extra—or one fewer—person can move the estimate," so computing the IF gives the sampling variance *without* resampling.

From Influence Functions to Inference

David McCoy PhD, MSc Division of Biostatistics, UC Berkeley Edward's Lifesciences

Definition and Simple Examples

Substitution Estimator for Causal Parameters: ATE

G-Comp Estimation of the ATE

Non-parametric Boostrap

Inverse Propensity of

Influence functions provide standard errors

$$\text{SE} = \sqrt{\frac{\sum_{i=1}^{n} \text{IF}_i^2}{n}}$$

Central Limit Theorem:

- Estimate converges to normal distribution
- Variance determined by influence function
- Valid for complex estimators (IPTW, Targeted Learning)

Practical advantages:

- Direct variance estimation
- No bootstrap required
- Works for any smooth estimator

Influence Function for IPTW

$$\mathrm{IF}_i = \frac{w_i A_i}{\bar{w}_1}\left(Y_i - \bar{Y}_1\right) - \frac{w_i(1-A_i)}{\bar{w}_0}\left(Y_i - \bar{Y}_0\right) + \hat{\psi}_{\mathrm{IPTW}}$$

- $\bar{w}_1 = \frac{1}{n}\sum_i w_i A_i$, $\bar{Y}_1 = \dfrac{\sum_i w_i A_i Y_i}{\sum_i w_i A_i}$ (analogous for $A = 0$).
- **Standard error**: $\widehat{\mathrm{SE}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} \mathrm{IF}_i^2}$.
- **HeartFix numbers** ($n = 1000$): $\hat{\psi}_{\mathrm{IPTW}} = -8$ pp; $\widehat{\mathrm{SE}} = 1.8$ pp; 95
- Interpretation — big weights on "unlikely" treated or untreated patients, but each person's IF still averages out to zero, giving a valid variance in one shot.

Let's Code It

Load the iptw.R file in Rstudio and let's look it over.
Now we will be estimating the same ATE on the same simulated HEARTFIX data but with IPTW instead of g-comp.

G-Computation vs. IPTW: Same Target, Different Tactics

David McCoy PhD, MSc
Division of Biostatistics, UC Berkeley
Edward's Lifesciences

Definition and Simple Examples

Substitution Estimator for Causal Parameters: ATE

G-Comp Estimation of the ATE

Non-parametric Boostrap

Inverse Propensity of

Aspect	**G-Computation**	**IPTW**
Part of likelihood	Models $P(Y \mid A, W)$ directly	Weights by $P(A \mid W)$; cancels confounding
Modeling focus	Outcome regression (Super-Learner)	Propensity score (SuperLearner)
Inference	Non-parametric bootstrap	Influence function (analytic SE)
Positivity issues	Low: extrapolates smoothly	High: extreme weights $\rightarrow$ large variance
Best when	- Complex estimator (IF hard to derive) - Strong outcome signal - Overlap concerns	- Good overlap - Need quick SE - IF is known/simple

Take-home

Both target the same counterfactual means ($[Y_1], [Y_0]$) but shift modeling burden to different parts of the likelihood.

IPTW Diagnostics: Propensity-Score & Weight Checks

David McCoy PhD, MSc
Division of Biostatistics, UC Berkeley
Edward's Lifesciences

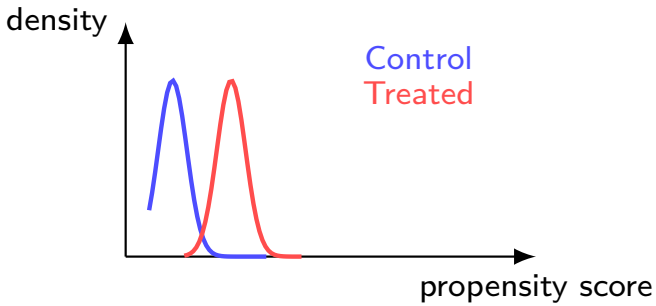
Definition and Simple Examples

Substitution Estimator for Causal Parameters: ATE

G-Comp Estimation of the ATE

Non-parametric Boostrap

Inverse Propensity of



Non-overlap when curves barely intersect

- **Overlap check**
  Plot PS distributions by treatment
  Little overlap $\Rightarrow$ unstable IPTW
- **Key metrics**
  - Effective sample size:
    $n_{\text{eff}} = (\sum w_i)^2 / \sum w_i^2$
  - Max / 99th percentile weight
  - Truncation range (e.g., 1–99%)
- **Rules of thumb**
  - $n_{\text{eff}} < 0.25n \Rightarrow$ trim
  - Max weight $> 10 \Rightarrow$ investigate
- **Poor diagnostics?**
  Trim, truncate, or use AIPW/TMLE

Stochastic Interventions: the Big Picture

→ **Beyond "treat everyone / treat no-one".**
A *stochastic intervention* replaces the observed treatment rule with a *new probability* of treatment,

$$g^\star(A = 1 \mid W) \;\text{ vs. }\; g_0(A = 1 \mid W) = P_0(A = 1 \mid W).$$

→ **Examples.**

- Vaccination campaign that raises uptake by 20 pp in every risk group.
- Policy that caps opioid prescriptions at 50
- "Nudge" changing the default from opt-in to opt-out.

→ **Target parameter.**
Counterfactual mean outcome under the new policy:

$$\psi(g^\star) \;=\; P_0\Big[\underbrace{P_0\big\{Y \mid A, W\big\}}_{\text{biology}} \,\Big|\, A \sim g^\star(\cdot \mid W)\Big].$$

→ **Why care?** Tells policy-makers the *population-level* impact of nudging probabilities—often more realistic than all-or-nothing scenarios.

Estimating $\psi(g^\star)$ with IPTW

David McCoy PhD, MSc Division of Biostatistics, UC Berkeley Edward's Lifesciences

Definition and Simple Examples

Substitution Estimator for Causal Parameters: ATE

G-Comp Estimation of the ATE

Non-parametric Boostrap

Inverse Propensity of

→ **Key trick:** re-weight each subject by the likelihood *ratio*

$$w_i = \frac{g^\star(A_i \mid W_i)}{\hat{g}(A_i \mid W_i)},$$

where $\hat{g}$ is any flexible propensity-score learner.

→ **Hájek estimator (mean with normalised weights)**

$$\hat{\psi}_{\text{shift}} = \frac{\sum_{i=1}^{n} w_i \, Y_i}{\sum_{i=1}^{n} w_i}.$$

→ **Variance in one line.** Influence function for a weighted mean $\Rightarrow$ $\widehat{\text{SE}} = \sqrt{\frac{1}{n}\sum_i (w_i(Y_i - \hat{\psi}))^2}$.

✎ **Up next:** jump into the `iptw_stochastic_shift.R` lab to:

1. fit $\hat{g}$ with `SuperLearner`;
2. apply a $+20$ pp shift;
3. compute $\hat{\psi}$, its IF-based SE and 95 % CI;
4. diagnose weights and overlap in code.

Summary

- Same method can be used for many parameters related to theoretical interventions at one time (point treatment cases), including:
  - → treatment rules,
  - → mediation impacts (direct and indirect effects),
  - → stochastically assigned interventions,
  - → etc.
- Substitution estimators are "relatively" intuitive.
- When estimating the regressions with parameteric models, can use the (nonparametric) bootstrap to get inference (or the delta-method).
- Though these estimators work when machine learning is used, the don't result in an estimator with predictable sampling distribution.

Summary

- Without modifications, one cannot get inference (e.g., confidence intervals).
- We need to understand the sampling distribution of our estimator - thus we need the influence function for a parameter.
- IPTW gets us there but is sensitivity to propensity estimator error
- That's why we need TMLE - a doubly robust estimator which uses all parts of the likelihood.

References

J.M. Robins.
Association, causation, and marginal structural models.
Synthese, 121:151–179, 1999.